SUNY College of Environmental Science and Forestry

Digital Commons @ ESF

Honors Theses

5-2013

Characterization by solution small angle X-ray scattering of the oligomeric state of structural genomics protein targets

Kimberly Basil

Follow this and additional works at: https://digitalcommons.esf.edu/honors

 Part of the Plant Sciences Commons

Recommended Citation

Basil, Kimberly, "Characterization by solution small angle X-ray scattering of the oligomeric state of structural genomics protein targets" (2013). *Honors Theses*. 12.
https://digitalcommons.esf.edu/honors/12

This Thesis is brought to you for free and open access by Digital Commons @ ESF. It has been accepted for inclusion in Honors Theses by an authorized administrator of Digital Commons @ ESF. For more information, please contact digitalcommons@esf.edu, cjkoons@esf.edu.

Characterization by solution small angle X-ray scattering of the oligomeric state of structural genomics protein targets

by

Kimberly Basil
Candidate for Bachelors of Science
Department of Environmental Science and Forestry
With Honors

May 2013

APPROVED

Thesis Project Advisor: Christopher Whipps, Ph.D.

Second Reader: Lee Newman, Ph.D

Honors Director: William M. Shields, Ph.D.

Date: April 17, 2013

Abstract

Solution small angle X-ray scattering (SAXS) was used to confirm the oligomeric state of several structural genomics protein targets from the New York Structural GenomiX Research Consortium. The oligomeric state is representative of the natively active form of the protein and it was determined to corroborate the Protein Data Bank (PDB) structure of a clone or related protein. PDBs are usually modeled from X-ray crystallography and the results of crystalline structures are not often representative of the correct biological assembly. Therefore, the analysis done here was essential to find the natural form. Using solution SAXS allowed us to look at each protein in its native state dissolved in a solution, rather than placed into a crystal array where it is unknown if interactions between each molecule are forced or natural. After obtaining a scattering pattern, several software programs from the ATSAS program suite for small-angle scattering data analysis were utilized to distinguish the quaternary structures of the proteins from monomers to tetramers. The final products were low-resolution three-dimensional structures accurate to the native form of the proteins.

Table of Contents

Acknowledgements

This project would not have been possible without the support of many people. I would like to thank my helpful mentor Dr. Marc Allaire for his expert guidance on this project and whose assistance was invaluable. Appreciation is owed to Dr. Lee Newman who introduced to me this huge opportunity at Brookhaven National Laboratory and who has been a consistent guiding presence, Dr. Christopher Whipps who has been an active and helpful advisor, and Dr. Bill Shields for encouraging me and providing many words of counsel since the honors seminar my first semester here. Sincere gratitude is also due to my family for their continued and loving support and to all the friends I have made here these four years. I would also like to convey thanks to Brookhaven National Laboratory and the Department of Energy for providing financial means and a laboratory facility.

Introduction

It is largely assumed in structural studies of biological macromolecules that finding the three-dimensional (3D) structure is necessary to understand how the macromolecules function. This paradigm has led to increasing growth in structural analysis studies and projects, especially in high-resolution structure determination of individual proteins using X-ray crystallography (XRC). This technique that has been widely used since the start of the Genomics Era when the amount of available genome sequences has increased rapidly[1]. However, XRC relies on the protein's crystalline structure, which can be a tedious and timely process to make and often displays molecules not accurate to their oligomeric state. During this technique, a protein crystal, a 3D array of ordered and highly packed protein molecules held together by noncovalent interactions, is created and modeled. Then, the smallest asymmetrical unit (ASU) of the crystal is deposited in the Protein Data Bank (PDB)[2]. The ASU, though the correct structure for one of the chains for the protein, may not represent the whole biological unit. For example, a monomeric protein may crystallize with multiple copies of the polypeptide chain in the ASU; but it may function as a dimer where it takes two ASU structures to complete its task in the biological system. The opposite may occur, as well, where a dimeric protein crystalizes as a single chain[3]. Therefore, the proposed quaternary state of the proteins the PDB has listed needs to be found using alternative techniques.

While there are theoretical approaches using machine-learning methods to predict this level of structure, it is widely accepted that the quaternary structure of proteins must be determined empirically[4]. The protein quaternary structure file

server and the Protein Interfaces, Surfaces and Assemblies method are two common techniques developed predict the biological state from its crystal structure, but errors arise in their outputs because it is hard to determine which interactions in the ASU are biological or forced[5]. Thus, it may be necessary to complete the description of the oligomeric state of proteins in the PDB without relying on the crystal structure.

Because some structural genomics projects are aiming to enhance PDB data with information about the protein's quaternary structure, they are exploiting solution small angle X-ray scattering (SAXS) as a quick and relatively simple technique to ascertain the overall shape of a protein without the limits of a crystal array. Four structural genomics protein targets sent from the New York Structural GenomiX Research Consortium (NYSGXRC) were examined for this experiment at the National Synchrotron Light Source (NSLS) on beamline X9 at Brookhaven National Laboratory (BNL). Using a synchrotron source for X-rays is useful to study weak scattering systems, typical of biological molecules, in very a short time[6]. The X-ray beam interacts with the sample and secondary waves scatter off of the molecules. These waves are detected as intensities (I) over a range of angles (ϑ) and converted to the scattering curve (I(q)) with use of the scattering vector amplitude (q), the wavelength of the radiation (λ) and the equation $q = 4\pi(\sin \vartheta)/ \lambda$[7]. These proteins, called by the PDB codes 3NF4, 3LKE, 3KFO and 3NF2, were associated with an already known crystal structure listed in the PDB archive. 3NF4 refers to the clone of acyl-CoA dehydrogenase from *Mycobacterium thermoresistibile*, while 3LKE is related to enoyl-CoA hydratase of *Bacillus halodurans*. 3KFO is a relative of the C-

terminal domain from the nuclear pore complex component NUP133 of *Saccharomyces cerevisiae*, and 3NF2 is related to the polyprenyl synthetase from *Streptomyces coelicolor.*

The goal of this experiment is to show that solution SAXS is a reliable tool to address the oligomeric state of proteins in the current wave of structural genomics projects. The hypothesis is that there are no discrepancies regarding the overall quaternary state from the data compiled from X9 to the already hypothesized biological model of the cloned or related proteins found in the PDB.

Experimental Methods and Materials

Each protein was dissolved in the buffer, 10mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid at the pH of 7.5, 150mM sodium chloride, 10mM Methionine and 10% glycerol. The concentrations of protein to buffer were the following: 3N4F at 11.5 mg/ml, 3LKE at 13.74 mg/ml, 3KFO at 10.5 mg/ml, and 3NF2 at 10.3 mg/ml. These solutions were sent directly from NYSGXRC to BNL. Individual solutions were sent through the X9 beam at the NSLS. The scattering data from the proteins was picked up by both a wide angle X-ray scattering detector and a SAXS detector providing a combined q range from 0.005 $Å^{-1}$ to 2.0 $Å^{-1}$. Only the SAXS q range was looked at, up to 0.25 $Å^{-1}$ because it is in this range that contains data of the quaternary structure. Software pyXS, developed at beamline X9, was used for data processing, including azimuthal averaging, averaging of a 2D image provided from the detector into a 1D intensity function, I(q), and background subtraction[8]. From here, components of the ATSAS software program were utilized

for data analysis. Analysis began with PRIMUS to manipulate the raw data[9]. In this program, the pure buffer signal was subtracted from the solution signal. The radius of gyration of the molecule (Rg) was then found using AutoRg by using the Guinier approximation, $\ln I(q) = \ln I(0) - (Rg^2/3)(q^2)$, in PRIMUS[10]. Subsequently, a Guinier plot was created to see if there was any aggregation in the sample. GNOM was then run to calculate the particle distance distribution function (P(r)), which indicates what conformation the protein is in, using the Fourier transform and from estimating the D_{max}, the maximum dimension of the molecule, from the Rg[11,12]. It also created the output file necessary to run the modeling programs DAMMIN and GASBOR. After the proper output from GNOM was made, *ab initio* protein shape determination using a dummy atom model was performed by DAMMIN ten times. A similar modeling program, GASBOR, was also used three times for 3NF2 to validate the results. DAMMIN used stimulated annealing, where the D_{max} set in GNOM places the constraints on the size of the protein and the program fills the properly sized sphere with densely packed smaller beads whose positions change places until a best fit model is found[13]. GASBOR also followed the parameters set by GNOM, but employed an ensemble of dummy residues to form a chain-compatible model[14]. Both these programs built slightly different models each time they ran because multiple structures can fit a single scattering curve. DAMAVER aligned all of the models, selected the most common one and constructed an averaged model[15]. The experimental scattering data was then fit to the known PDB data using CRYSOL[16]. CRYSOL compared the two curves in a single plot and was the main tool that was utilized to see the oligomeric state of the proteins. If the two curves overlapped in

the SAXS region, then the known PDB information contained the correct quaternary structure. On the other hand, if they did not line up, the XRC data were modified by removing duplicated data from the PDB until the curves fit. The final step was to use SASREF, which modeled the quaternary structure of the proteins with use of both the solution scattering data and the known and possibly modified PDB data[17].

Results and Discussion

The first protein, 3N4F, was found to have an Rg of 19.97 and minimal aggregation, which was avoided in data analysis by omitting the first 20 points of the scattering data. The D_{max} was set to 60 and DAMAVER provided a model (Fig. 1). CRYSOL was used to compare the scattering data and the PDB data. The PDB data of the experimental protein's clone contained coordinates for four chains in the protein, but the experimental data implied a monomer (Fig. 2), where there is a poor fit between the two curves. After three chains were removed from the PDB data, CRYSOL was run again and the error between the two curves was decreased (Fig. 3). The SASREF image (Fig. 4), produced for 3N4F with the modified PDB and experimental data also corroborated the monomeric result for the quaternary structure of the protein.

The next protein, 3LKE, had an Rg of 27.94 and a D_{max} set to 84. The first 18 points were omitted in GNOM. This protein had a trimer for its oliomgeric state and DAMAVER was used to model 3LKE (Fig. 5). Unlike the CRYSOL results from 3N4F, the first run in CRYSOL showed a good fit between the experimental and known data (Fig. 6). To further support this, two chains were removed from the PDB data

and the program was run again. This produced a poor fit between the two data sets (Fig. 7). The SASREF result for 3LKE was a trimer that was nearly identical to the PDB that can be downloaded from the PBD database (Fig. 8).

The third protein, 3KFO, had an Rg of 19.65 and the first 13 points from the experimental data were removed in GNOM. The modeled structures from both DAMAVER (Fig. 9) and SASREF (Fig. 10) are shown. The D_{max} was set to 60. These results were similar to 3LKE such that the XCR PDB data were a match to the SAXS data for its quaternary structure; both produced results demonstrated a monomer.

The last protein modeled was 3NF2. The Rg was 27.71 and in GNOM the first 12 points were left out. The D_{max} was set to 84 and the oliogomeric state of the protein was found out to be a dimer, but the XCR data showed a monomer. The results mimicked 3N4F's two CRYSOL runs where the first had a poor fit between the two curves and the second output was much better. In this case however, in the second CRYSOL run, an extra chain was added to the XRC data instead of removing chains. The DAMAVER (Fig. 11) and SASREF output (Fig. 12) both indicated the protein was a dimer.

In the case of each protein, the experimental results from beamline X9 were verified by the proposed biological model in the PDB, found using alternative methods relying on the crystal array. This confirms if the oligomeric state of a protein has yet to be predicted, solution SAXS is a technique that can be utilized without developing its ASU.

It is important to note that although the CRYSOL outputs contain the most crucial information about confirming the oligomeric state results, the finished

SASREF models provide a more complete version of the protein than what is already deposited in the PDB. As previously mentioned in the introduction, the model of the biological unit provided from methods depending on the crystal may contain errors from possible forced attractions. As a result, some predictions are not indicative of the actual biologically active quaternary structure. In one study, it was found that the error rate of models found using the protein quaternary structure server is estimated to be around 16 percent[18]. The SASREF models, in this experiment, provide an answer to this error rate where the solution SAXS data delivers information that creates a prediction less altered by the possible unnatural attractions found in crystal arrays and consequently, a more accurate model.

Conclusion

This experiment confirms the reliability of solution SAXS as a technique used to confirm the structure of a protein, and therefore provides positive evidence for the previously stated hypothesis. When most structural genomics protein projects rely on XRC and related methods to confirm structure, SAXS allows scientists to address a fundamental piece of the puzzle not often known or if known, incorrectly established, the oligomeric state. The functional form of each protein discussed was found and modeled with little prior knowledge of them: 3N4F, a monomer; 3LKE, a trimer; 3KFO, a monomer and 3NF2, a dimer. These data can be used to further biological studies on these four specific proteins and for other closely related biological macromolecules. Even though there is a strong connection between protein structure and function, future progress in this field should focus on

increasing the methods used to structure the abundance of sequenced molecules, as well as refining the data and creating a level of standardization for SAXS to support these data.

Works Cited

[1]P. Bernadó, E. Mylonas, M.V. Petoukhov, M. Blackledge and D.I. Svergun, *J Am Chem. Soc.* **129** 5656-5664 (2007)

[2] G. Rhodes, *Crystallography made Crystal Clear: A Guide for Users of Macromolecular models,* (Academic Press, new York, 1993), pp.389

[3,18]E. Levy, *Structure.* **15** 1364-1367 (2007)

[4,5]R. Garian, *Bioinformatics.* **17** 551-556 (2001)

[6]J.Y. Bottero, D. Tchoubar, J.M. Cases, and F. Flessinger, *J Phys Chem.* **86** 3667-3673 (1982)

[7,12]M. Allaire, *SAXS/WAXS Applications*, Brookhaven National Laboratory Beamline X9 Workshop Handout (Unpublished)

[8]Y.K. Gupta, L. Yang, S. Chan, J.C. Samuelson, S. Xu, A.K. Aggarwal, *J Mol Bio.* **420** 261-268 (2012)

[9]P.V. Konarev, V.V. Volkov, A.V. Sokolova, M.H.J. Koch and D.I. Svergun, *J Appl Cryst.* **36** 1277-1282 (2003)

[10]M.V. Petoukhov, P.V. Konarev, A.G. Kikhney and Svergun, *J Appl Cryst.* **40** s223-s228 (2007)

[11]D.I. Svergun, *J Appl Cryst.* **25** 495-503 (1992)

[13]D.I. Svergun, *Biophys J.* **25** 2879-2886 (1999)

[14]D.I. Svergun, M.V. Petoukhov and M.H.J. Koch *Biophys J.* **80** 2946-2953 (2001)

[15]V.V. Volkov and D.I. Svergun, *J Appl Cryst.* **36** 860-864 (2003)

[16]D.I. Svergun, C. Barberato and M.H.J. Koch, *J Appl Cryst*. **28** 768-773 (1995)

[17]M.V. Petoukhov and D.I. Svergun, *Biophys J.* **89** 1237-1250 (2005)

Appendices

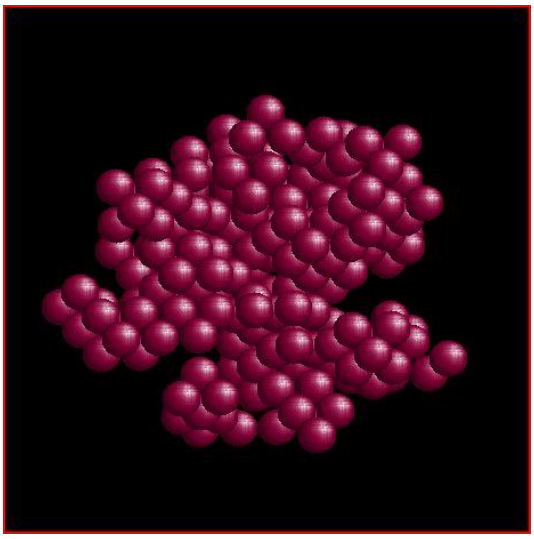

Figure 1. Image of DAMAVER result of 3N4F after ten runs of DAMMIN modeling.

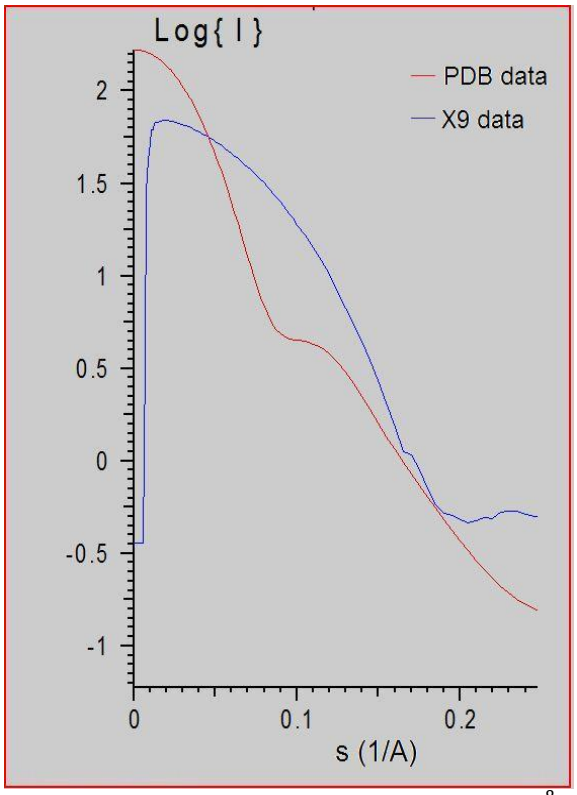

Figure 2. CRYSOL output cut to 0.25 (1/Å), labeled as s. In this case, s is comparative to *q*. The blue line, representing the experimental data, does not line up with the red line, representing the given data.

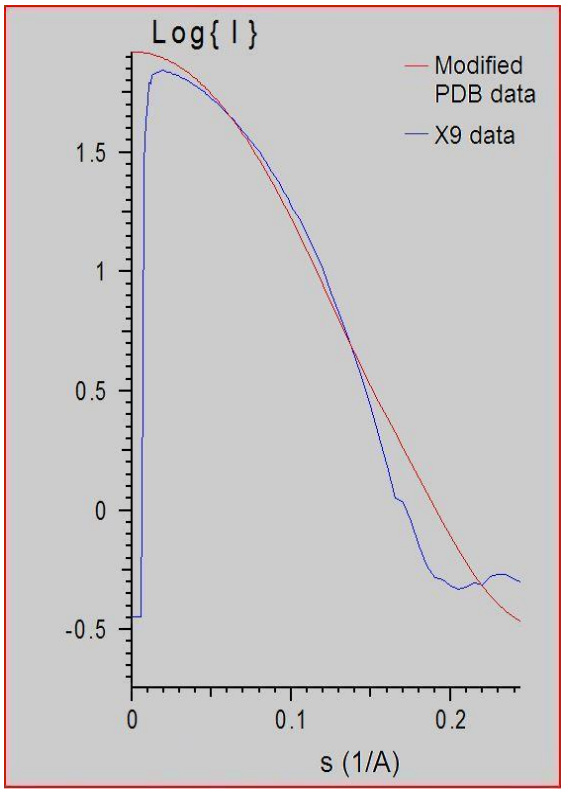

Figure 3. CRYSOL output of one chain from the crystalline PDB data compared to the scattering data of 3N4F.

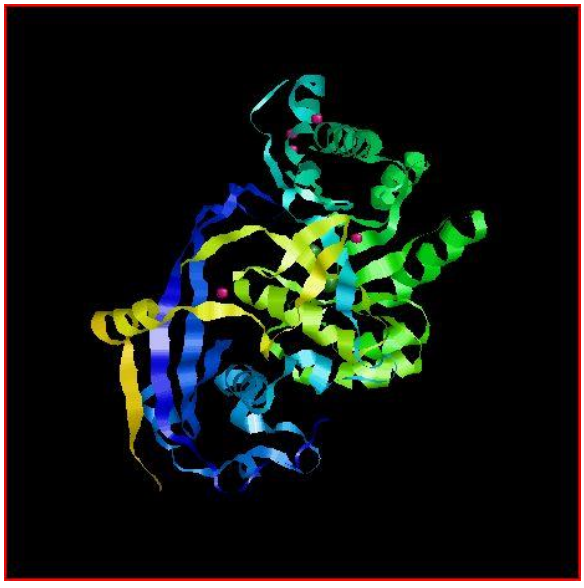
Figure 4. SASREF result of the monomer 3N4F.

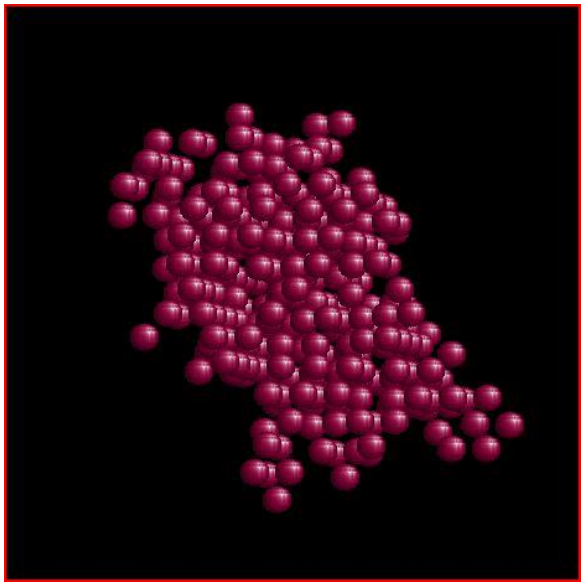

Figure 5. DAMAVER result of 3LKE after ten runs in DAMMIN.

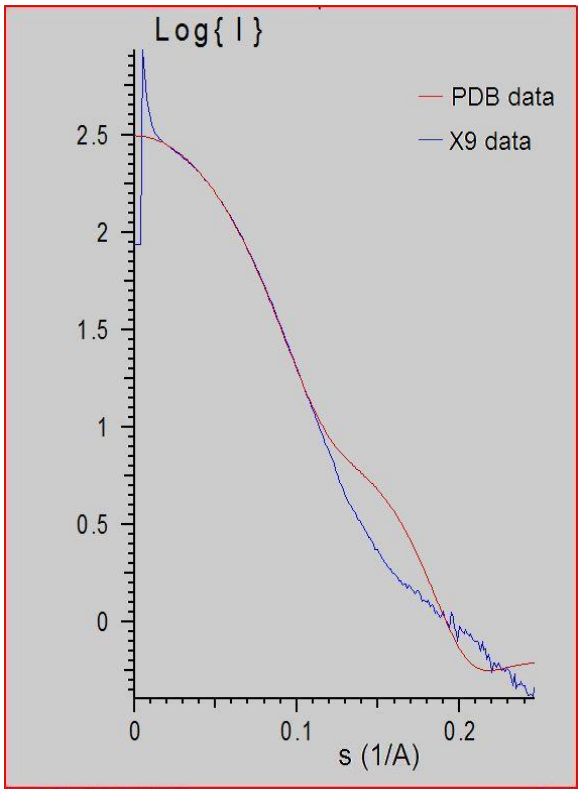

Figure 6. 3LKE XCR data compared to SAXS data in CRYSOL. This confirms the PDB information is representative of the biological unit with three chains.

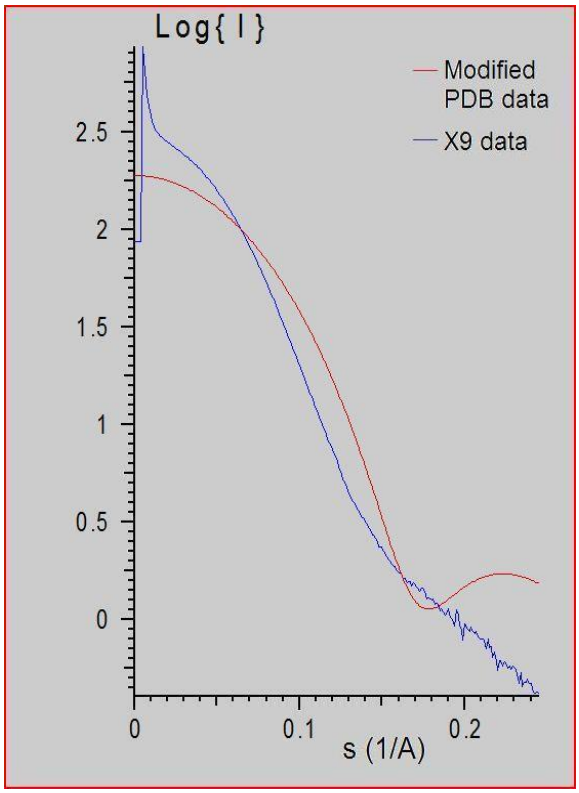

Figure 7. Modified 3LKE XCR data to one chain compared to the SAXS trimer data run in CRYSOL. The original XCR data has a better fit.

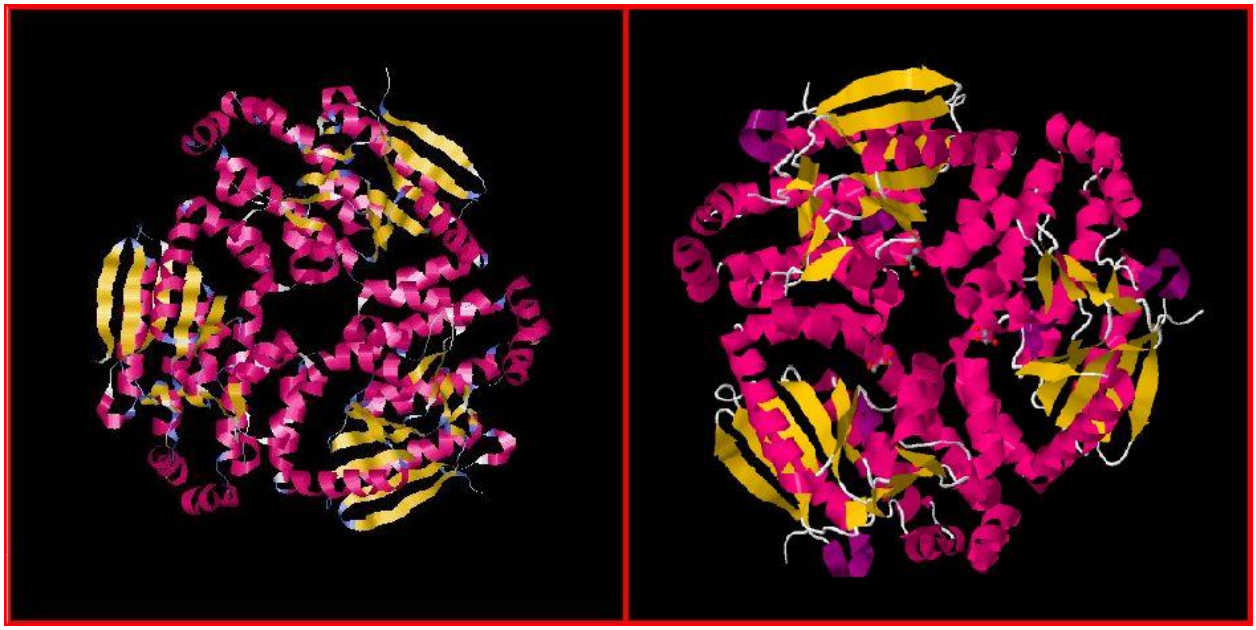
Figure 8. SASREF result of 3LKE, on the right, next to the downloaded XCR PDB image, on the left where both images are very similar.

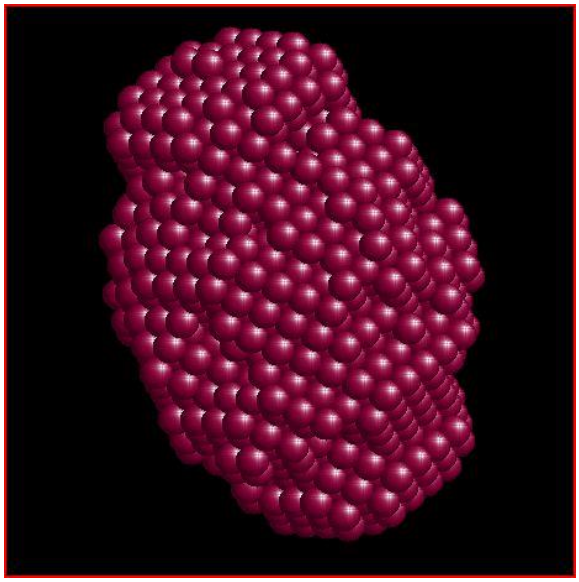

Figure 9. DAMAVER image produced from ten DAMMIN runs of 3KFO. The protein is a monomer.

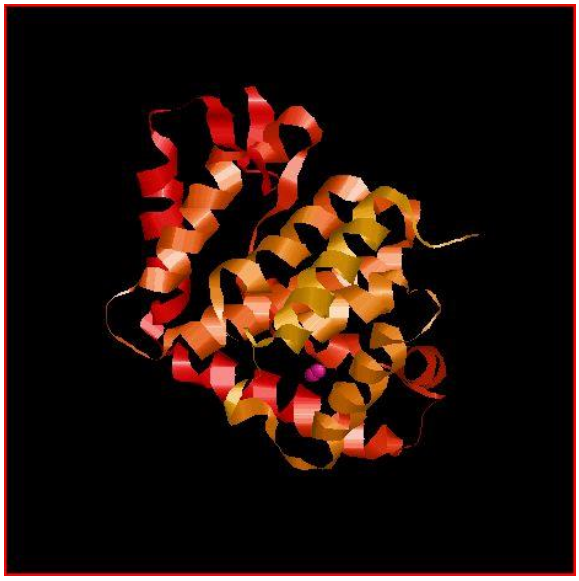

Figure 10. Monomeric SASREF result of 3KFO.

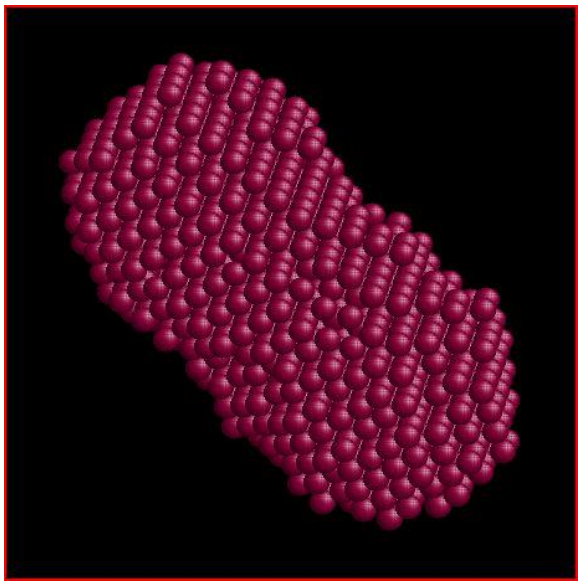

Figure 11. The average shape of 3NF2 found with DAMAVER of ten DAMMIN runs and three GASBOR runs. It contains the area for two protein chains.

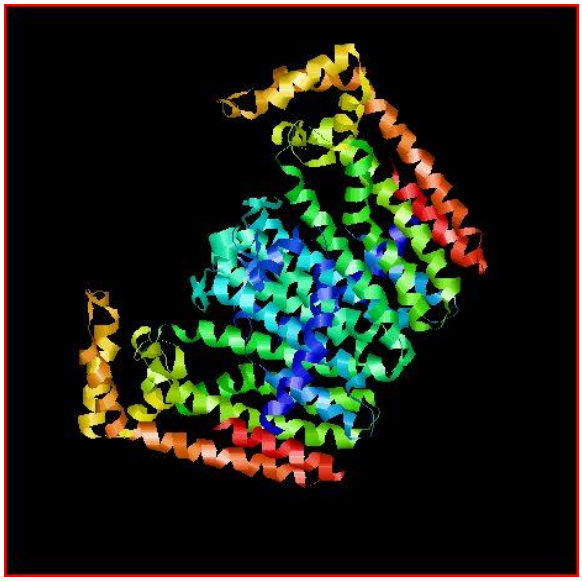

Figure 12. SASREF result of the dimer of 3NF2.